Air pollution Analysis with a PFCM Clustering Algorithm Applied in a Real Database of Salamanca (Mexico)

B. Ojeda-Magaña[1], M. G. Cortina-Januchs[2], J. M. Barrón-Adame[3], J. Quintanilla-Domínguez[2], W. Hernandez [4], A. Vega-Corona[3], R. Ruelas[1] and D. Andina[2]

[1]University of Guadalajara
Projects Engineering Department (DIP-CUCEI).
[2]Technical University of Madrid
Group for Automation in Signals and Communications (GASC).
[3]University of Guanajuato
Faculty of Engineering Mechanics, Electrical and Electronics. Salamanca, México.
[4]Technical University of Madrid
Department of Circuits and Systems (ICS)
Email: benojed@hotmail.com; andina@gc.ssr.upm.es.

***Abstract*— Over the last ten years, Salamanca has been considered among the most polluted cities in México. Nowadays, there is an Automatic Environmental Monitoring Network (AEMN) which measures air pollutants (Sulphur Dioxide (SO_2), Particular Matter (PM_10), Ozone (O_3), etc.), as well as environmental variables (wind speed, wind direction, temperature, and relative humidity), and it takes a sample of the variables every minute. The AEM Network is mainly based on three monitoring stations located at Cruz Roja, DIF, and Nativitas. In this work, we use the PFCM (Possibilistic Fuzzy c Means) clustering algorithm as a mean to get a combined measure, from the three stations, looking to provide a tool for better management of contingencies in the city, such that local or general action can be taken in the city according to the pollution level given by each station and the combined measure. Besides, we also performed an analysis of correlation between pollution and environmental variables. The results show a significative correlation between pollutant concentrations and some environmental variables. So, the combined measure and the correlations can be used for the establishment of general contingency thresholds.**

I. Introduction

AAir pollution is one of the most important environmental problems in developed and undeveloped countries. Pollution is caused by both natural and man-made sources, and it may vary greatly from one region to another according to the geography, demography, climate, and topography of these ones. For example, pollutant concentration decreases significantly if the urban area has special topography or large rainy season [2]. Major man-made sources of air pollution include: industries, transportation, agriculture, power generation, and unplanned urban areas [1].

Sulphur Dioxide (SO_2), and Particular Matter (PM_{10}) are the air pollutants with the highest concentration in Salamanca, where three monitoring stations have been installed in order to know the level of air pollution; the measure records of each monitoring station are handled separately. Actually, an environmental contingency alarm is activated when daily average pollutant concentration, in a single monitoring station, exceeds a established threshold.

In this work, we propose to apply the PFCM (Possibilistic Fuzzy c Means) clustering algorithm in order to get a combined measure from data of the three monitoring stations, such that local environmental contingency alarms can be taken, according to the pollutant concentration reported by each monitoring station, and general (or city) environmental contingency alarms that will depend on the levels provided by the combined measure. So, the PFCM algorithm is used as a way to find the prototypes of the patterns that represent the relation between SO_2 and PM_{10} air pollutants. In the relation analysis we use the records from January 2007.

Once prototypes have been estimated, we realize a comparison between pollution averages of each monitoring station and the prototypes. In the analysis we use a data set from January to December 2007. Analysis includes pollutants concentrations, SO_2 and PM_{10}, and meteorological variables, wind speed, wind direction, temperature, and relative humidity.

We also analyze the impact of meteorological variables on the dispersion of pollutants through the calculus of correlation coefficients. The correlation analysis is very simple and of great interest looking for improved decision making in environmental programs. Only data of Nativitas station is used in the correlation analysis.

This paper is organized as follow: In Section II we explain the features of the area under study, and explain the air pollution problem in Salamanca. In Section III we introduce the PFCM (*Possibilistic Fuzzy c Means*) clustering algorithm and

978-1-4244-5697-0/10/$25.00 ©2010 IEEE

the correlation coefficients. Section IV presents the obtained results. And finally, in Section V we present our conclusions.

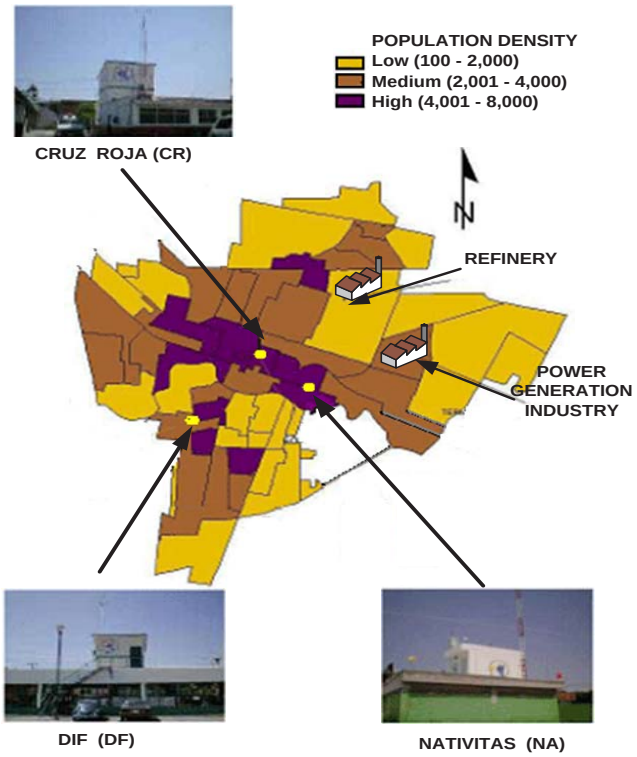


Fig. 1. Location of monitoring stations in the city of Salamanca.

II. Study Case

Salamanca is located in the state of Guanajuato, Mexico, and it has an approximate population of 234,000 inhabitants [4]. The city is 340 km northwest from Mexico City, with coordinates $20°34'22"$ North latitude, and $101°11'39"$ West longitude. It is located on a valley surrounded by the *Sierra Codornices*, where there are elevations with an average height of 2,000 meters Above Mean Sea Level (AMSL).

Salamanca has been one of the Mexican cities with more important industrial development in the last fifty years. The Refinery and Power Generation Industries have settled down in the fifty and seventy decades, respectively. These industries constitute the main and most important energy source for local, regional and national economy. However, the population's increase, the vehicular park, the industry, the refinery and thermoelectric activities, as well as orography and climatic characteristics have propitiated the increment in SO_2 and PM_{10} concentrations [5]; the orography difficult the dispersion of pollutants by the wind, which produces the worst pollutant concentrations. The SO_2 emissions are bigger than those emitted in the Metropolitan area of Mexico City or Guadalajara, even when these ones have a bigger population than the city of Salamanca [6]. The Orography hinder the dispersion of the worst pollutants by winds.

Sulfur dioxide is produced fundamentally by the combustion of fossil fuels, and it has the energy generation sector as the main source of pollution. That is, the industrial sector generates 99.3 % of this pollutant, and only an approximate percentage of 0.06 % is generated by the transport sector. Particles produced by electric power generation represent 29 % of the total emissions, it follows the vehicular traffic in the roads without paving with 27 %, next the agriculture burns with 17 %, the transport sector with 10 %, and the remaining 17 % is emitted by other sub-sectors.

Authorities of the city have made important efforts to measure and record on concentrations of pollutants [7]. In 1999 the Air Quality Monitoring Patronage (AQMP) was formed. Since then the AQMP has been in charge of running the Automatic Environmental Monitoring Network (AEMN), and disseminate information. This information is validated by the Institute of Ecology (IE), which constantly analyzes the levels of pollutants [5]. The AEMN consists of three fixed and one mobile stations. The fixed stations are: Cruz Roja, Nativitas, and DIF.

The fixed stations cover approximately 80 % of the urban area while the mobile station covers the remaining 20 %. Fig. 1 illustrates the location of the three fixed stations. Each station has the necessary instrumentation to automatically track every minute concentration of pollutants and meteorological variables. Table I contains a sample of the concentration of pollutants and meteorological variables in each of the three fixed stations.

Pollutants

	Cruz Roja	Nativitas	DIF
Ozone (O_3)	√	√	√
Sulfur Dioxide(SO_2)	√	√	√
Carbon Monoxide (CO)	√	√	√
Nitrogen Dioxide (NO_x)	√	√	√
Particulate Matter less than 10 micrometer in diameter (PM_{10})		√	√

Meteorological variables

	Cruz Roja	Nativitas	DIF
Wind Direction (WD)	√	√	√
Wind speed (WS)	√	√	√
Temperature (T)		√	√
Relative Humidity (RH)		√	√
Barometric Pressure (BP)		√	√
Solar Radiation (SR)		√	√

√ Measured

TABLE I

Pollutants concentrations and meteorological variables recorder in the monitoring stations

III. Clustering Algorithms

The objective of the fuzzy clustering algorithms is to find an internal structure in a numerical data set into *n* different subgroups, where the members of each subgroup have a high similarity with its prototype (centroid, cluster center,

signature, template, code vector) and a high dissimilarity with the prototypes of the other subgroups. This justifies the existence of each one of the subgroups [8].

A simplified representation of a numerical data set into *n* subgroups, help us to get a better comprehension and knowledge of the data set [9]. Besides, the particional clustering algorithms (hard, fuzzy, probabilistic or possibilistic) provide, after a learning process, a set of prototypes as the most representative elements of each subgroups.

Ruspini was the first one to use fuzzy sets for clustering [10]. After that, Dunn [11] developed in 1973 the first fuzzy clustering algorithm, named Fuzzy c-Means (FCM), with a parameter of fuzziness *m* equal to 2. Later on Bezdek [12] generalized this algorithm. The FCM is an algorithm where the membership degree of each point to each fuzzy set A_i is calculated according to its prototype. The sum of all the membership degrees of each individual point to all the fuzzy sets must be equal to one.

Krishnapuram and Keller [13] developed the Possibilistic c-Means (PCM) clustering algorithm, where the principal characteristic is the relaxation of the restriction that gives the relative typicality property of the FCM. The PCM provides a similarity degree between data points and each one of the prototypes, value known as absolute typicality or simply typicality [14]. So, the nearest points to a prototype are identified as typical, whereas the furthest points as atypical, and noise [15] [17].

A. PFCM clustering algorithm

Pal et al. [14] have proposed to use the membership degrees as well as the typicality values, looking for a better clustering algorithm. They called it *Fuzzy Possibilistic c-Means* (FPCM). However, the sum equal to one of the typicality values for each point was the origin of a problem, particularly when the algorithm uses a lot of data. In order to avoid this problem, Pal *et al* [16] proposed to relax this constraint and they developed the PFCM clustering algorithm, where the function to be optimized is given by (1)

$$J_{pfcm}(\mathbf{Z};\mathbf{U},\mathbf{T},\mathbf{V}) = \sum_{i=1}^{c}\sum_{k=1}^{N}(a\mu_{ik}^{m}+bt_{ik}^{\eta})\times\|z_k-v_i\|^2 + \sum_{i=1}^{c}\gamma_i\sum_{k=1}^{N}(1-t_{ik})^{\eta}, \quad (1)$$

and subject to the constraints $\sum_{i=1}^{c}\mu_{ik} = 1\forall k$; $0 \le \mu_{ik}, t_{ik} \le 1$ and the constants $a > 0$, $b > 0$, $m > 1$ and $\eta > 1$. The parameters *a* and *b* define a relative importance between the membership degrees and the typicality values. The parameter μ_{ik} in (1) has the same meaning as in the FCM. The same happens for the t_{ik} values with respect to the PCM algorithm.

Theorem PFCM [16]: If $D_{ikA} = \|z_k - v_i\| > 0$, for every i, $k, m, \eta > 1$, and Z contains at least *c* different patterns, then $(U,T,V) \in M_{fcm} \times M_{pcm} \times \Re^p$ and J_{pfcm} can be minimized if and only if

$$\mu_{ik} = \left(\sum_{j=1}^{c}\left(\frac{D_{ikA_i}}{D_{jkA_i}}\right)^{2/(m-1)}\right)^{-1} \quad (2)$$
$$1 \le i \le c; \quad 1 \le k \le n$$

$$t_{ik} = \frac{1}{1+\left(\frac{b}{\gamma_i}D_{ik_{A_i}}^2\right)^{1/(\eta-1)}} \quad (3)$$
$$1 \le i \le c; \quad 1 \le k \le n$$

$$v_i = \sum_{k=1}^{N}(a\mu_{ik}^{m}+bt_{ik}^{m})z_k \Big/ \sum_{k=1}^{N}(a\mu_{ik}^{m}+bt_{ik}^{m}), \quad (4)$$
$$1 \le i \le c.$$

The membership degrees are calculated with equation (2), the typicality values with (3) and for the prototypes the equation (4) is used.

B. PFCM clustering algorithm in the AEMN

As it is known, in the partition clustering algorithms is necessary a minimum of two groups. However, in our problem we only have one group, this group is formed by patterns [SO_2;PM_{10}] pollutant concentrations. Therefore, we propose a synthetic cloud of patterns with the following covariance matrix and vector of centers:

$\sum_1 = \begin{bmatrix} 400 & 0 \\ 0 & 400 \end{bmatrix}$, $\begin{bmatrix} v_1 \end{bmatrix} = \begin{bmatrix} 100 & -600 \end{bmatrix}$.

In this case, the number of patterns (4320) is the same in the synthetic cloud and the pollutant concentration.

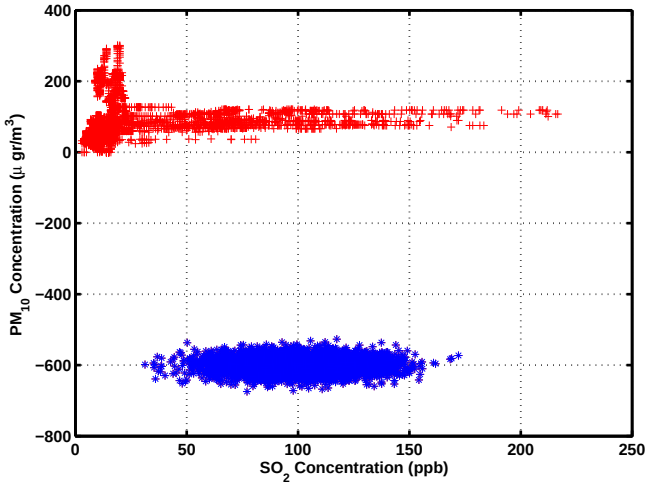


Fig. 2. Air pollution and synthetic cloud patterns.

Fig. 2 shows clearly the synthetic cloud (located in the lower part) and the pollutant concentration patterns (located in the

superior part). Once the groups are identified, we apply the PFCM clustering algorithm.

C. Correlation Coefficient

The correlation coefficient r (also called Pearson's product moment correlation after Karl Pearson [18]) is used to determine the strength and direction of the relationship between two variables. This form of correlation requires that both variables are normally distributed, interval or ratio variables. The correlation coefficient is calculated by eq.(5):

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum {x_i}^2) - (\sum x_i)^2}\sqrt{n(\sum {y_i}^2) - (\sum y_i)^2}} \quad (5)$$

where n is the number of data points. The numerical values of correlation coefficient range from +1 to -1. If two variables move exactly together, the value of the correlation coefficient is 1. This indicates perfect positive correlation. If two variables move exactly opposite to each other, the value of the correlation coefficient is -1. Low numerical values indicate little relationship between two variables, such as -0.10 or +0.15 indicate little relationship between on two variable.

IV. Results

Fig. 3 shows the distribution of pollutant patterns [SO_2;PM_{10}] at the three monitoring stations (CR, DF and NA). The mesh in Fig. 3 corresponds to the thresholds established by the program to improve the air quality in Salamanca (*ProAire*) [5]. Thresholds are Pre-contingency, Phase-I contingency and Phase-II contingency. For example, for SO_2 concentrations equal to or bigger than 145 *ppb* and smaller than 225 *ppb* (average per day), a level of environmental pre-contingency is declared. Therefore the spaces between lines in the mesh represent the levels of environmental contingency for SO_2 and PM_{10} concentrations. In Fig. 3 each symbol

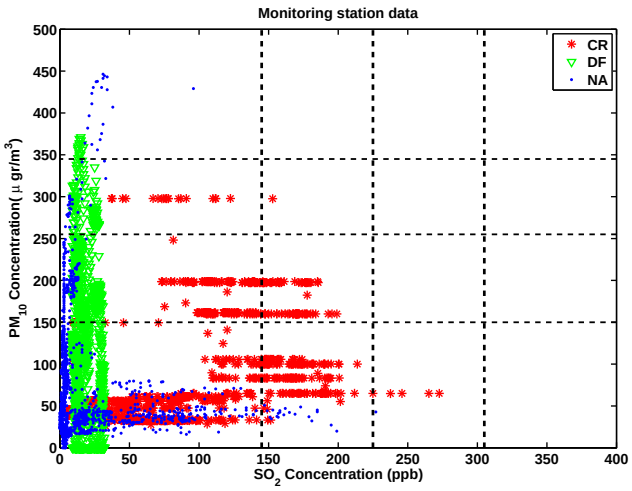


Fig. 3. Monitoring Network per minute.

(*, • and ▽) represent the pollutant patterns at each monitoring station. At Nativitas monitoring station we observe that the highest PM_{10} and SO_2 pollutant concentrations are not present at the same time. On other hand, at the Cruz Roja monitoring station we observe that either SO_2 or PM_{10} pollutant concentrations are highest. At the DIF monitoring station we observe the highest PM_{10} concentrations in the AEMN network.

Our proposal in this work is to apply the PFCM clustering algorithm to the AEMN in Salamanca. The main proposal is to integrate the pollutant measures from the three monitoring stations.

The PFCM initial parameters (a, b, m and η) are very important in order to reduce the outlier effects in the pattern prototypes. Pal *et al*, in [16] recommend of b parameter value larger than the a parameter value in order to reduce the mentioned effects. On the other hand, a small value for η and a value greater than 1 for m are recommended. nevertheless, choosing a too high of a value of m reduces the effect of membership of data to the clusters, and the algorithm behaves as a simple PCM.

Taking into account the previous recommendations, the initial parameters for the PFCM clustering algorithm were set as follows: $a = 1$, $b = 5$, $m = 2$ and $\eta = 2$. We present the found prototypes (a and b) in Fig. 4.

In Fig. 4(a) the daily averages of SO_2 concentrations are presented for each monitoring station together with the corresponding prototypes. We observed also that Cruz Roja monitoring station receives the highest emissions of SO_2 concentrations: this is due to it bieng located very near to the refinery. The prototypes in this case were very low in comparison with the observed SO_2 concentrations, because only one station observed high SO_2 concentrations (Cruz Roja). According with analyzed patterns the emitted pollutant is only measured by the Cruz Roja monitoring station (see Fig. 4).

Fig. 4(b) shows the daily averages of PM_{10} concentrations and result prototypes. In this case, the observed averages are very similar at the three monitoring stations. The PM_{10} pollutant dispersion is more uniform then the SO_2 pollutant dispersion in the city. Table II shows the correlation results among SO_2 and PM_{10} pollutants and the meteorological variables. The database used in the correlation analysis correspond to year 2004 of Nativitas. This period was taking because contains more meteorological registrations. The obtained results of the SO_2 correlation coefficient show a high positive correlation between SO_2 pollutant and Wind Speed, also a high and negative correlation between SO_2 pollutant and Wind Direction is observed. The other meteorological variables have not impact. For the PM_{10} pollutant, the meteorological variable with more impact is the Relative Humidity. We observe, when the Relative Humidity increases the pollutant concentration decreases. The PM_{10} particles are caught and fall to the ground during rain.

V. Conclusions

Nowadays, there is a program to improve the air quality in the city of Salamanca, Mexico. Besides, this program

	SO_2	PM_{10}
SO_2	1	0.0731
PM_{10}	0.0731	1
WS	0.4756	-0.1385
WD	-0.6151	0.1478
T	-0.0329	-0.0007
RH	-0.0322	-0.4416
BP	0.1462	0.1806
SR	-0.021	-0.1207

TABLE II
CORRELATION COEFFICIENT BETWEEN POLLUTANT CONCENTRATION AND METEOROLOGICAL VARIABLES.

has established thresholds for several levels of contingencies depending on the SO_2 and PM_{10} pollutant concentrations. However, a particular level of contingency for the city is declared taking into account the highest pollutant concentration provided by one of the three monitoring stations. For example, if a pollutant concentration exceeds a given threshold in a single monitoring station, the alarm of contingency applies to the whole city. This value is normally provided by the Cruz Roja station, due to its proximity to the refinery and power generation industries.

Looking for local and general contingency levels in the city, we have proposed to estimate a set of prototypes such that they can represent a calculated measure of pollutant concentrations according to the values measured in the three fixed stations. In such a way, a local alarm of contingency can be activated in the area of impact of the pollution depending on each station, and a general alarm of contingency according to the values provided by the prototypes. Nevertheless, the last case requires adjusting the thresholds, as the actual values would be only used for local contingency because they depend on the measured values of pollutant concentrations, and the general contingency requires thresholds as a function of calculated values.

In this work we use the PFCM clustering algorithm for the estimation of the prototypes for the general contingency alarm. The prototypes correspond to the nearest points to those of data. That means, they are mean points that depend on a criterion of minimum distance to the values measured in each monitoring station. These are more logical values for the activation of general contingency alarms, although it is necessary to clearly identify the area of impact of the pollution depending on the measured values at each station.

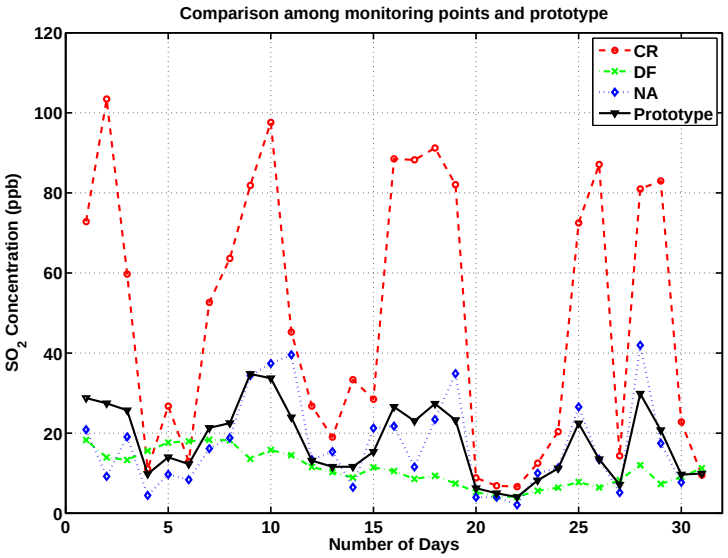
Besides, we have studied the correlation among pollutant concentrations and meteorological variables, because if they are correlated they must be taken into account when defining the local area of contingency alarm, as well as for the prediction of the duration of each particular contingency. We have used data of the Nativitas monitoring station for that. From the correlation analysis results, wind speed and wind direction show a significative degree of correlation with the SO_2 concentration, whereas relative humidity show a similar correlation with the PM_{10} pollutant concentration. As the meteorological variables have an important impact on the concentration of pollutants, they will be used in a near future for the definition of the local and general contingency levels.

ACKNOWLEDGEMENTS

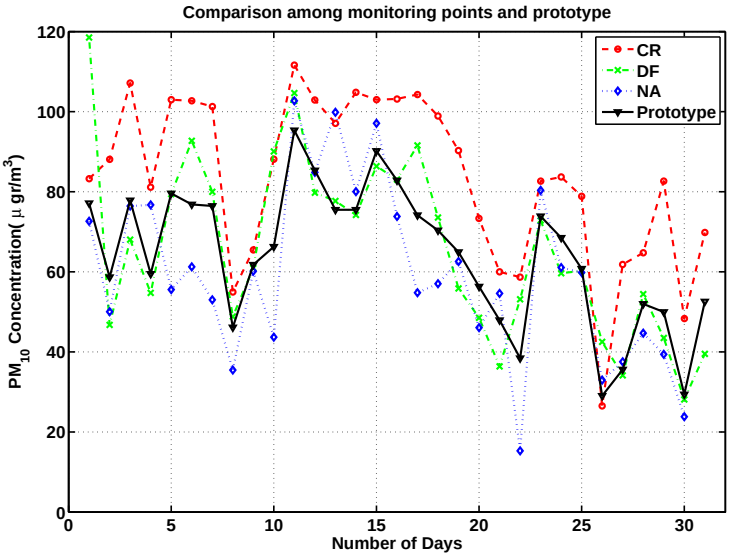
This work has been partially supported by the Ministry of Science and Innovation (MICINN) of Spain under the research project TEC2007-63121; the Computational Intelligence Laboratory (LABINCO) of Guanajuato University, Consejo Nacional de Ciencia y Tecnología (CONACyT) and Secretaría de Educación Pública (SEP), Mexico.

REFERENCES

[1] Jes Fenger: “Air pollution in the last 50 years - From local to global”. *Journal of Atmospheric Environment*, vol. 43, pp. 13-22, 2009
[2] M.B. Celik, I. Kadi: “The relation between meteorological factors and pollutants concentration in Karabuk City”. *G.U. Journal of science*, vol. 20, pp. 87-95, 2007
[3] (In Spanish) Instituto de Ecología del Estado de Guanajuato: “Programa para mejorar la calidad del aire en Salamanca”. Calle Aldana Núm.12, Col. Pueblito de Rocha, 36040 Guanajuato, Gto., Segunda Edición. Abril 2004.
[4] National Institute of Geography and Statistics. “Population and Housing Census 2 (2005)”. *www.inegi.org.mx.*
[5] (In Spanish) Instituto de Ecología del Estado de Guanajuato: “Programa para mejorar la calidad del aire en Salamanca”. Calle Aldana Núm.12, Col. Pueblito de Rocha, 36040 Guanajuato, Gto., Segunda Edición. Abril 2004.
[6] M. G. Cortina-Januchs, J. M. Barron-Adame, A. Vega-Corona, D. Andina.: “revision of Industrial SO2 Pollutant Concentration Applying ANNs”. *th. IEEE International Conference on Industrial Informatics, INDIN*, 2009,pp. 510-515, DOI:10.1109/INDIN.2009.5195856.
[7] (In Spanish) A. Zamarripa, A. Sainez: “Medio Ambiente: Caso Salamanca”, *Instituto de Investigación Legistativa*, H. Congreso del Estado de Guanajuato, LX legislatura. 2007
[8] D. Andina and P. D. Truong, *Computational intelligence for engineering and manufacturing*, Springer Verlag, 2007.
[9] J. M. Barron-Adame, J. A. Herrera Delgado, M.G. Cortina-Januchs, D. Andina, A. Vega-Corona: “Air Pollutant Level Estimation Applying a Self-organizing Neural Network”. *Proceedings of the 2nd international work-conference on Nature Inspired Problem-Solving Methods in Knowledge Engineering. IWINAC ’07*, pp. 599-607, 2007.
[10] E. Ruspini. “Numerical method for fuzzz clustering”. *Inf. Sci. 2*, pages pp. 319–350, 1970.
[11] J.C. Dunn. “A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters”. *Journal of Cybernetics*, Vol 3,pp. 32–57, 1973.
[12] J. C. Bezdek. *Pattern Recognition With Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.
[13] R. Krishnapuram and J. Keller. “A possibilistic approach to clustering”. *International Conference on Fuzzy Systems*, Vol. 1(2): 98–110, 1993.
[14] N.R. Pal, S.K. Pal and J.C. Bezdek. “A mixed c-means clustering model”. In *IEEE International Conference on Fuzzy Systems, Spain*, pp. 11-21, 1997.
[15] B. Ojeda-Magaña, R. Ruelas, F.S. Buendía-Buendía and D.Andina. “A Greater Knowledge Extraction Coded as Fuzzy Rules and Based on the Fuzzy and Typicality Degrees of the GKPFCM Clustering Algorithm”. *In Intelligent Automation and Soft Computing*, Vol. 15(4): 555–571, 2009.
[16] N.R. Pal, S.K. Pal, J.M. Keller and J.C. Bezdek. “A possibilitic fuzzy c-means clustering algorithm”. *IEEE Transactions on Fuzzy Systems*, 13(4):517–530, 2005.
[17] B. Ojeda-Magaña, J. Quintanílla-Dominguez, R. Ruelas, and D. Andina, “Images sub-segmentation with the pfcm clustering algorithm,” *Proceedings of The 7th. International Conference on Industrial Informatics (INDIN 09)*, pp. 499–503, 2009.
[18] P. Pérez, A. Trier, J. Reyes: “Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile”. *Atmospheric Environment*, vol. 9, pp. 1189 - 1196, 2000.



(a) SO_2



(b) PM_{10}

Fig. 4. Comparison between air pollutant averages and estimated prototypes.